Multivariate Plug-in Bandwidth for Local Linear Regression

By LIJIAN YANG and ROLF TSCHERNIG*
Michigan State University, East Lansing, USA Humboldt-Universität, Berlin, Germany

March 9, 1998

Abstract

Optimal bandwidths for local polynomial regression usually involve functionals of the derivatives of the unknown regression function. In the multivariate case, estimates of these functionals are not readily available, primarily because estimating multivariate derivatives is complicated. In this paper, an estimator of multivariate second derivative is obtained via local quadratic regression with cross terms left out. This estimator has the optimal rate of convergence but is simpler and uses a lot less computing time than the full local quadratic estimator. Using this as a pilot estimator, an estimator of the integrated squared Laplacian of a multivariate regression function is obtained which leads to a plug-in formula of the optimal bandwidth for multivariate local linear regression. This bandwidth has good theoretical properties as well as satisfactory performance in our simulation study. It is also recommended for variable selection methods.

Abbreviated Title. Multivariate Plug-in Bandwidth

Keywords: ASYMPTOTIC OPTIMALITY, BANDWIDTH SELECTION, FUNCTIONAL ESTIMATION, LAPLACIAN, LOCAL QUADRATIC REGRESSION, SECOND DERIVATIVES, WITHIN BIAS TRADE-OFF

1. INTRODUCTION

Nonparametric estimation in general requires little a priori knowledge on the functions to be estimated. The estimation results, however, depend crucially on the bandwidth choice. While choosing too large a bandwidth may introduce a large bias, selecting too small a bandwidth may cause large estimation variance. An asymptotically optimal bandwidth usually exists and can be obtained through a bias-variance trade off. Such an optimal bandwidth in general involves functionals of the unknown underlying functions, and selection of the optimal bandwidth for various function estimation problems has always been a challenge. Bandwidth selection methods with good theoretical properties and practical performance exist for univariate density estimation (see, for example, Jones, Marron and Sheather, 1996a,b, Cheng 1997, Grund and Polzehl, 1997) and univariate local least squares regression (see, for example, Fan and Gijbels 1995, Ruppert, Sheather and Wand,

* *Address for Correspondence:* Institut for Statistics and Econometrics, Humboldt University, Spandauer Str. 1, D-10178 Berlin, E-mail: rolf@wiwi.hu-berlin.de; *Acknowledgments*: The authors thank Rong Chen and Michael Neumann for their many helpful discussions. This work was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" of the Deutsche Forschungsgemeinschaft, at Humboldt-Universität zu Berlin.

1995). Wand and Jones (1993) contains some excellent analysis of bandwidth selection for bivariate density estimation. Ruppert (1997) proposes empirical-bias bandwidth selector (EBBS) which is applicable in multivariate setting. However, the theoretical properties of EBBS is not known and its practical performance has only been studied in the univariate setting. As pointed out by Ruppert (1997), the difficulty in obtaining a reliable multivariate data-driven bandwidth is essentially due to the complexity of estimating higher order multivariate derivatives. Our paper attempts to address the multivariate bandwidth issue via a simple algorithm to estimate second order multivariate derivatives.

To be precise, consider a multivariate regression model

$$Y = f(\mathbf{X}) + g(\mathbf{X})\epsilon \tag{1.1}$$

where Y is a scalar dependent variable, $\mathbf{X} = (X_1, X_2, ..., X_d)$ is a vector of explanatory variables, ϵ is independent of $\mathbf{X}$, $E\epsilon = 0$ and $var(\epsilon) = 1$. Let $(\mathbf{X}_i, Y_i)$, $i = 1, 2, ..., n$ be an i.i.d. sample. Then the local linear estimator of f at a given point $\mathbf{x} = (x_1, x_2, ..., x_d)$ is obtained by doing a first order Taylor expansion of the function f at point $\mathbf{x}$ for all the data points $\mathbf{X}_i$ and solve a least squares problem locally weighted by kernels: i.e., the estimator $\widehat{f}(\mathbf{x})$ is the first element c of the vector

$$\left\{c, (c_\alpha)_{1\le\alpha\le d}\right\}$$

that minimizes

$$\sum_{i=1}^{n}\left\{Y_i - c - \sum_{\alpha=1}^{d} c_\alpha(X_{i\alpha} - x_\alpha)\right\}^2 K_h(\mathbf{X}_i - \mathbf{x}) \tag{1.2}$$

where K is a symmetric, compactly supported, univariate probability kernel (so that K is nonnegative and $\int K(u) = 1$), and

$$K_h(\mathbf{X}_i - \mathbf{x}) = \frac{1}{h^d}\prod_{\alpha=1}^{d} K\left(\frac{X_{i\alpha} - x_\alpha}{h}\right). \tag{1.3}$$

Denoting by $p(\mathbf{x})$ the density of $\mathbf{X}$, the mean integrated squared error (MISE) $E\int\left\{\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right\}^2 p(\mathbf{x})d\mathbf{x}$ is a function of the bandwidth h, and the h that minimizes this error is called the optimal bandwidth. An asymptotic optimal bandwidth in this setting is given by (see e.g. Tschernig and Yang (1997))

$$h_{opt} = \{d||K||_2^{2d}B(g)n^{-1}C(f)^{-1}\sigma_K^{-4}\}^{1/(d+4)} \tag{1.4}$$

in which $\sigma_K^2 = \int u^2K(u)du$, $||K||_2^2 = \int K^2(u)du$ and

$$B(g) = \int g^2(\mathbf{x})d\mathbf{x} \tag{1.5}$$

$$C(f) = \int\left\{\mathrm{Tr}\,\nabla^2 f(\mathbf{x})\right\}^2 p(\mathbf{x})d\mathbf{x} = \int\left\{\sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x})\right\}^2 p(\mathbf{x})d\mathbf{x}. \tag{1.6}$$

where $f_{\lambda\lambda}(\mathbf{x})$ denotes the second derivative of $f(\mathbf{x})$ with respect to the λ-th variable x_λ. The integration is always taken over the support of the density $p(\mathbf{x})$.

If the integral of $g^2(\mathbf{x})$ is infinite over the support of $p(\mathbf{x})$, one can always use in formula (1.5) and (1.6) a weight function with compact support, which is equivalent to screening off large observations, as done in Tjøstheim and Austead (1994).

The main difficulty for using (1.4) is the estimation of $C(f)$, the integrated squared Laplacian of f. To estimate $f_{\lambda\lambda}(\mathbf{x})$, one has to do a local quadratic regression, i.e., a second order Taylor expansion of $f(\mathbf{x})$. The local quadratic estimator of the second derivative $f_{\lambda\lambda}(\mathbf{x})$ is 2 times the $\alpha\alpha$-th element $c_{\alpha\alpha}$ of the vector

$$\left\{c, (c_\alpha)_{1\le\alpha\le d}, (c_{\alpha\alpha})_{1\le\alpha\le d}, (c_{\alpha\beta})_{1\le\alpha<\beta\le d}\right\}$$

that minimizes

$$\sum_{i=1}^{n}\left\{Y_i - c - \sum_{\alpha=1}^{d} c_\alpha(X_{i\alpha}-x_\alpha) - \sum_{\alpha=1}^{d} c_{\alpha\alpha}(X_{i\alpha}-x_\alpha)^2 - \sum_{1\le\alpha<\beta\le d} c_{\alpha\beta}(X_{i\alpha}-x_\alpha)(X_{i\beta}-x_\beta)\right\}^2 K_h(\mathbf{X}_i-\mathbf{x}). \tag{1.7}$$

A full Taylor expansion like that in equation (1.7) is unnecessary if one only needs to estimate $f_{\lambda\lambda}(\mathbf{x})$ for $C(f)$. In Section 3, it is shown that it suffices to solve a local least squares problem based on a partial expansion, ignoring all cross terms in the local quadratic expansion problem. This simplifies and speeds up the computation considerably. It will be shown that despite this facilitation the bias and the variance are of the same rates as keeping the cross terms in. The cost of leaving out the cross terms is essentially a more complicated bias formula (5.6) when estimating $f_{\lambda\lambda}$. It is also worth pointing out that integrating over $f_{\lambda\lambda}$ weakens the "curse of dimensionality" since the estimation of $C(f)$ achieves a convergence rate of $O(n^{-2/(d+6)})$ which is faster than the rate of $O(n^{-2/(d+8)})$ for estimating $f_{\lambda\lambda}$.

The paper is organized as follows. In the next section, we define a plug-in bandwidth selector for estimating $f(\mathbf{x})$, and give its asymptotic properties. It is called plug-in as it attempts to approximate h_{opt} by plugging in estimates of the unknown quantities $B(g)$ and $C(f)$ in formula (1.4). In Section 3, we describe in detail the asymptotic properties of the estimator of $C(f)$, and also of the partial local quadratic estimator of $f_{\lambda\lambda}(\mathbf{x})$ used for estimating $C(f)$. In Section 4, we present results from our simulation study. The results of this paper are also useful in variable selection procedures, such as developed in Tschernig and Yang (1997). All proofs are contained in the Appendix.

2. A PLUG-IN BANDWIDTH

We had already commented that the optimal bandwidth h_{opt} given in formula (1.4) contains unknown quantities $B(g)$ and $C(f)$ and hence cannot be directly used in the estimation of $f(\mathbf{x})$. A quick substitute is the simple rule-of-thumb bandwidth (Silverman, 1986, page 87), $h_{ROT} = h_S(d+2)$ where

$$h_S(k) = \sqrt{\widehat{\mathrm{var}}(\mathbf{X}_i)}\,\{4/k\}^{1/(k+2)}\, n^{-1/(k+2)} \tag{2.1}$$

in which

$$\widehat{\mathrm{var}}(\mathbf{X}_i) = \left\{ \prod_{\alpha=1}^{d} \widehat{\mathrm{var}}(X_{i\alpha}) \right\}^{1/d},$$

$$\widehat{\mathrm{var}}(X_{i\alpha}) = \frac{1}{n-1} \left\{ \sum_{i=1}^{n} X_{i\alpha}^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_{i\alpha} \right)^2 \right\}, \alpha = 1, 2, ..., d.$$

While $h_S(d+2)$ may look rather simple, it does have the same order of $n^{-1/(d+4)}$ as h_{opt} and is therefore a reasonable substitute of the latter. The problem is that it is not of the optimal ratio, i.e., one does not have $h_{ROT}/h_{opt} \to 1$ as $n \to \infty$. With some abuse, we are using the h_{ROT} designed for density estimation because there does not yet exist a simple rule-of-thumb bandwidth for multivariate regression. Although such a bandwidth exists in the univariate case (Fan and Gijbels, 1995), its multivariate generalization would be more complicated.

We propose a plug-in bandwidth

$$\widehat{h}_{opt} = \{d||K||_2^{2d} \widehat{B}(g) n^{-1} \widehat{C}(f)^{-1} \sigma_K^{-4}\}^{1/(d+4)} \tag{2.2}$$

in which $\widehat{B}(g)$ and $\widehat{C}(f)$ are estimates of $B(g)$ and $C(f)$, respectively. Here, we use

$$\widehat{B}(g) = \sum_{i=1}^{n} \left\{ Y_i - \widetilde{f}(\mathbf{X}_i) \right\}^2 / \widetilde{p}(\mathbf{X}_i)$$

with a local linear estimator $\widetilde{f}(\mathbf{x})$ and a kernel density estimator $\widetilde{p}(\mathbf{x})$, both with bandwidth $h_S(d+2)$. Using standard results such as contained in Wand and Jones (1995) or Fan and Gijbels (1996), one has $\widehat{B}(g)/B(g) = 1 + O_p(n^{-2/(d+4)})$ as $n \to \infty$.

The definition of $\widehat{C}(f)$ is given in (3.13) and its asymptotic properties given in Corollary 3.2, both in the next section. In particular, $\widehat{C}(f)/C(f) = 1 + O_p(n^{-2/(d+6)})$ as $n \to \infty$. Based on the asymptotics of $\widehat{B}(g)$ and $\widehat{C}(f)$, one has the following

Theorem 2.1 *Under Assumptions (A1) to (A4), for $n \to \infty$, the bandwidth defined in (2.2) is asymptotically optimal. In particular,*

$$\left(\frac{\widehat{h}_{opt} - h_{opt}}{h_{opt}} \right) = O_p(n^{-2/(d+6)}).$$

Since $\widehat{h}_{opt}$ is a consistent substitute for the optimal bandwidth h_{opt}, it can be shown that $\widehat{h}_{opt}$ performs asymptotically similarly to h_{opt} in terms of MISE since $\widehat{f}(\mathbf{x})$ using $\widehat{h}_{opt}$ becomes close to $\widehat{f}(\mathbf{x})$ with h_{opt}. See Neumann (1995) for details. $\widehat{h}_{opt}$ will on average work better than the naive h_{ROT} since the latter is an inconsistent bandwidth estimator. Simulation results in Section 4 strongly support these expectations.

3. LOCAL QUADRATIC ESTIMATION

The estimator $\widehat{f}_{\lambda\lambda}(\mathbf{x})$ obtained in Section 1 by solving the least squares problem (1.7) is given by

$$\widehat{f}_{\lambda\lambda}(\mathbf{x}) = 2e_{d+\lambda}^T \left(Z^T W Z\right)^{-1} Z^T W Y, \tag{3.1}$$

where $Y = (Y_i)_{n\times 1}$,

$$W = \operatorname{diag}\left\{\frac{1}{n}K_h(\mathbf{X}_i - \mathbf{x})\right\}_{i=1}^{n}, \tag{3.2}$$

$$Z = \left[1, \{X_{i\alpha} - x_\alpha\}_{1\le\alpha\le d}, \left\{(X_{i\alpha} - x_\alpha)^2\right\}_{1\le\alpha\le d}, \{(X_{i\alpha} - x_\alpha)(X_{i\beta} - x_\beta)\}_{1\le\alpha<\beta\le d}\right]_{i=1}^{n} \tag{3.3}$$

is a $n \times (d+1)(d+2)/2$ matrix and

$$e_\lambda \text{ is a } (d+1)(d+2)/2 \text{ vector of zeros whose } (\lambda+1)\text{-element is 1,} \tag{3.4}$$

see Ruppert and Wand (1994).

The complex formation of the matrix Z makes the explicit mathematical derivation of $\left(Z^T W Z\right)^{-1}$ rather difficult, and Z's size means that computing of (3.1) may be costly. Hence in this section we propose a way to reduce Z to a much simpler formation and a smaller size.

Since we are not interested in mixed derivatives $\frac{\partial^2}{\partial x_\alpha \partial x_\beta} f(\mathbf{x})$, the terms $c_{\alpha\beta}(X_{i\alpha} - x_\alpha)(X_{i\beta} - x_\beta)$ in (3.1) are not necessary. Thus one can alternatively define $\widehat{f}_{\lambda\lambda}(\mathbf{x})$ to be 2 times the $\alpha\alpha$-th element of the vector

$$\left[c, \{c_\alpha\}_{1\le\alpha\le d}, \{c_{\alpha\alpha}\}_{1\le\alpha\le d}\right]$$

that minimizes

$$\sum_{i=1}^{n}\left\{Y_i - c - \sum_{\alpha=1}^{d} c_\alpha(X_{i\alpha} - x_\alpha) - \sum_{\alpha=1}^{d} c_{\alpha\alpha}(X_{i\alpha} - x_\alpha)^2\right\}^2 K_h(\mathbf{X}_i - \mathbf{x}), \tag{3.5}$$

which is a reduced least squares problem, and equivalent to using a submatrix of the matrix Z in (3.3) in place of Z in the formula (3.1). Without creating new notations, we denote from now on this submatrix as

$$Z = \left[1, \{X_{i\alpha} - x_\alpha\}_{1\le\alpha\le d}, \left\{(X_{i\alpha} - x_\alpha)^2\right\}_{1\le\alpha\le d}\right]_{i=1}^{n}, \tag{3.6}$$

and now define e_λ as a $(2d+1)$ vector of zeros whose $(\lambda+1)$-element is 1. Our new estimator $\widehat{f}_{\lambda\lambda}(\mathbf{x})$ is then

$$\widehat{f}_{\lambda\lambda}(\mathbf{x}) = 2e_{d+\lambda}^T \left(Z^T W Z\right)^{-1} Z^T W Y. \tag{3.7}$$

In what follows, one denotes for any compact supported function L

$$\mu_r(L) = \int_{-\infty}^{\infty} u^r L(u)du,$$

hence in particular $\mu_2(K) = \sigma_K^2$. We write $\mu_4(K) = \kappa\sigma_K^4$ where κ denotes the kurtosis of the kernel K, which is always > 1. For the derivative estimation, we denote by $f_{\alpha\beta\gamma}(\mathbf{x})$ the derivative $\frac{\partial^3}{\partial x_\alpha \partial x_\beta \partial x_\gamma} f(\mathbf{x})$, etc.

Theorem 3.1 *Under assumptions (A1)-(A3) in the Appendix, for* $\lambda = 1, ..., d$, *as* $h \to 0$ *and* $nh^{d+4} \to \infty$, *one has*

$$\sqrt{nh^{d+4}} \left\{ \widehat{f}_{\lambda\lambda}(\mathbf{x}) - f_{\lambda\lambda}(\mathbf{x}) - b_{\lambda\lambda}(\mathbf{x}) h^2 \right\} \longrightarrow N(0, \sigma^2_{\lambda\lambda}(\mathbf{x})) \tag{3.8}$$

where

$$\sigma^2_{\lambda\lambda}(\mathbf{x}) = \frac{\mu_4(K^2) \left\| K \right\|_2^{2(d-1)} g^2(\mathbf{x})}{\sigma_K^8 (\kappa - 1)^2 p(\mathbf{x})} \tag{3.9}$$

in which $p(\mathbf{x})$ *is the density of* $\mathbf{X}$. *The formula for* $b_{\lambda\lambda}(\mathbf{x})$ *is in equation (5.6) of the Appendix.*

Remember that our goal is to estimate the optimal bandwidth (1.4) which requires the estimation of the functional $C(f) = \int \left\{ \sum_{\lambda=1}^d f_{\lambda\lambda}(\mathbf{x}) \right\}^2 p(\mathbf{x}) d\mathbf{x}$. We now propose to estimate $C(f)$ with $\widehat{C}(f,h) = \int \left\{ \sum_{\lambda=1}^d \widehat{f}_{\lambda\lambda}(\mathbf{x}) \right\}^2 p(\mathbf{x}) d\mathbf{x}$, where the partial local quadratic estimator $\widehat{f}_{\lambda\lambda}(\mathbf{x})$ is given in (3.7), $\lambda = 1, 2, ..., d$. Our next theorem gives the asymptotic property of $\widehat{C}(f,h)$. In the following, we denote $K^*(u) = K(u)u^2$, $K * K^*$ is the convolution between K and K^*, $(K * K^*)(t) = \int K(t-u) K^*(u) du$, which equals $\int K(t+u) K^*(u) du$ because K^* is symmetric; and $K^{(2)}$ is the convolution of K with itself.

Theorem 3.2 *Under assumptions (A1)-(A4) in the Appendix, as* $h \to 0$ *and* $nh^{d+4} \to \infty$, *one has*

$$\widehat{C}(f,h) = C(f) + 2h^2 \int \sum_{\lambda=1}^d f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^d b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

$$+ \frac{\int g^2(\mathbf{x}) d\mathbf{x} \left\{ d\mu_4(K^2) \left\| K \right\|_2^{2(d-1)} + d(d-1) \mu_2^2(K^2) \left\| K \right\|_2^{2(d-2)} \right\}}{\sigma_K^8 (\kappa - 1)^2 n h^{d+4}} + \zeta + O_p\left(h^4\right) + o_p\left(\frac{1}{n\sqrt{h^{d+8}}} \right)$$

in which

$$n\sqrt{h^{d+8}} \zeta \xrightarrow{D} N\left\{ 0, \sigma^2(g, K, d) \right\}$$

where

$$\sigma^2(g, K, d) = \frac{2 \int g^4(\mathbf{x}) d\mathbf{x}}{\sigma_K^{16} (\kappa - 1)^4} \int F^2(\mathbf{t}) d\mathbf{t}$$

and

$$F(\mathbf{t}_1) = \sum_{1 \le \lambda \ne \mu \le d} (K^* * K)(t_{1\mu}) (K^* * K)(t_{1\lambda}) \left\{ \prod_{\gamma \ne \lambda, \mu} K^{(2)}(t_{1\gamma}) \right\} + \sum_{1 \le \lambda \le d} K^{*(2)}(t_{1\lambda}) \left\{ \prod_{\gamma \ne \lambda} K^{(2)}(t_{1\gamma}) \right\}.$$

Note from the theorem that $\widehat{C}(f,h)$ has an asymptotic standard deviation that is of smaller order than one of the bias terms: $\left(n\sqrt{h^{d+8}} \right)^{-1} = o\left((nh^{d+4})^{-1} \right)$ and the two bias terms point to a trade-off if both are positive, or cancellation if the h^2 term happens to be negative. If the h^2 term happens to be zero, then higher order terms need to be considered, but that is beyond the scope of this paper. Therefore we obtain the following results for the optimal bandwidth for estimating $C(f)$.

Corollary 3.1 *Under the same assumptions as in Theorem 3.2, to estimate* $C(f)$ *by* $\widehat{C}(f,h)$
(1) if $\int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} < 0$, *then*

$$h_{C(f),opt} = \left\{ -\frac{\int g^2(\mathbf{x}) d\mathbf{x} \left\{ d\mu_4(K^2) \left\| K \right\|_2^{2(d-1)} + d(d-1)\mu_2^2(K^2) \left\| K \right\|_2^{2(d-2)} \right\}}{2 \int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \sigma_K^8 (\kappa - 1)^2 n} \right\}^{\frac{1}{d+6}} \tag{3.10}$$

is the optimal bandwidth, in which case the asymptotic bias is of order $h_{C(f),opt}^4 = O(n^{-4/(d+6)})$ *and standard error of order* $n\sqrt{h^{d+8}} = O(n^{-(d+4)/2(d+6)})$;
(2) if $\int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} > 0$, *then*

$$h_{C(f),opt} = \left\{ (d+4) \frac{\int g^2(\mathbf{x}) d\mathbf{x} \left\{ d\mu_4(K^2) \left\| K \right\|_2^{2(d-1)} + d(d-1)\mu_2^2(K^2) \left\| K \right\|_2^{2(d-2)} \right\}}{4 \int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \sigma_K^8 (\kappa - 1)^2 n} \right\}^{\frac{1}{d+6}} \tag{3.11}$$

is the optimal bandwidth, in which case the asymptotic bias is of order $h_{C(f),opt}^2 = O(n^{-2/(d+6)})$ *and standard error of order* $n\sqrt{h^{d+8}} = O(n^{-(d+4)/2(d+6)})$.

Now observe that all the terms in the denominator are either known or have to be estimated anyway like $B(g) = \int g^2(\mathbf{x}) d\mathbf{x}$. Estimating, however, the term $\int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ in the denominator or knowing its sign can be extremely difficult, thus it is not feasible to know which one of (1) and (2) is the case and compute a plug-in bandwidth $\widehat{h}_{C(f),opt}$ accordingly. However, Corollary 3.1 shows that the optimal rate for $h_{C(f),opt}$ is independent of the sign of $\int \sum_{\lambda=1}^{d} f_{\lambda\lambda}(\mathbf{x}) \sum_{\lambda=1}^{d} b_{\lambda\lambda}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ and always $n^{-1/(d+6)}$. Since there does not exist any rule-of-thumb bandwidth we suggest to use the bandwidth

$$h_C = 3h_S(d+4) = 3\sqrt{\widehat{\mathrm{var}}(X_t)} \left\{ 4/(d+4) \right\}^{1/(d+6)} n^{-1/(d+6)} \tag{3.12}$$

which has the correct order $1/(d+6)$. This choice of h_C is admittedly crude due to the need for a simple working solution. By Corollary 3.1 $h_{C(f),opt}$ minimizes the absolute value of the asymptotic bias and hence deviating from $h_{C(f),opt}$ in either direction always increases the bias. By Theorem 3.2, however, increasing the bandwidth always leads to a reduction in the variance of $\widehat{C}(f)$. It is therefore preferable to overestimate $h_{C(f),opt}$ than to underestimate it, which is our reason for preferring $3h_S(d+4)$ to $h_S(d+4)$. It turns out to produce satisfactory results in our simulation study in section 4.

We therefore define

$$\widehat{C}(f) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{\lambda=1}^{d} \widehat{f}_{\lambda\lambda}(\mathbf{X}_i) \right\}^2 = \int \left\{ \sum_{\lambda=1}^{d} \widehat{f}_{\lambda\lambda}(\mathbf{x}) \right\}^2 p(\mathbf{x}) d\mathbf{x} + O_p(n^{-1/2}), \tag{3.13}$$

where the partial local quadratic estimator $\widehat{f}_{\lambda\lambda}(\mathbf{x})$ is given in (3.7), $\lambda = 1, 2, ..., d$ and using bandwidth h_C. Definition (3.13) and Theorem 3.2 yield

Corollary 3.2 *Under the Assumptions (A1) to (A4), for $n \to \infty$, the $\widehat{C}(f)$ defined by (3.13) has the following consistency property:*

$$\left\{ \frac{\widehat{C}(f) - C(f)}{C(f)} \right\} = O_p(n^{-2/(d+6)}).$$

We want to add a comment here that the results of this section can be generalized to autoregressive time series under conditions that lead to geometric mixing property, see, for instance, Härdle, Tsybakov and Yang (1997) for such conditions. The pilot estimator $\widehat{C}(f)$ was used in the plug-in formula of a data-driven asymptotic final prediction error (AFPE) for nonlinear time series lag selection developed by Tschernig and Yang (1997).

4. SIMULATION RESULTS

In this section we investigate the finite sample performance of the simple rule-of-thumb bandwidth (2.1) and the plug-in bandwidth (2.2) for the local linear estimation of the regression function $f(\mathbf{x})$. These results are compared with those which are obtained by using the known asymptotic optimal bandwidth (1.4). For each pseudo data set generated, we compute the estimated MISE (EMISE)

$$\frac{1}{n} \sum_{i=1}^{n} (\widehat{f}_h(\mathbf{X}_i) - f(\mathbf{X}_i))^2 \tag{4.1}$$

for $h = h_{opt}, \widehat{h}_{opt}, h_{ROT}$. In total we consider four different regression functions in the regression model

$$Y = f(\mathbf{X}) + \epsilon, \quad \epsilon \sim N(0, 0.01) :$$

- Model 1: $f(\mathbf{X}) = (X_1^2 + X_2^2 + X_3^2 + X_4^2)/4$
- Model 2: $f(\mathbf{X}) = X_1^2 X_2^2 + X_3^2 X_4^2$
- Model 3: $f(\mathbf{X}) = X_1^3/3 + (X_2 - 0.3)^2 X_3^2 X_4^2$
- Model 4: $f(\mathbf{X}) = X_1^2 X_2^2 + \sin(\pi X_3/2)/4 + \cos(\pi X_4/2)/4$

For all models we use a random design matrix X whose elements are independently drawn from a uniform distribution in the interval $[-0.5, 0.5]^4$ and two sample sizes: 100 and 500. In all cases we conduct 100 replications. Note that since $f(\mathbf{x})$, $g(\mathbf{x})$ and $p(\mathbf{x})$ are explicitly given, we can compute h_{opt} by (1.4). We are comparing the h_{ROT} designed for density estimation with our $\widehat{h}_{opt}$ simply due to the nonexistence of a rule-of-thumb bandwidth for multivariate regression, as we commented in Section 2.

All the results are shown in Figures 1 and 2. The results for each model are depicted in one column of graphs. For Model 1, Figure 1 (a) shows the densities of EMISE for the optimal bandwidth h_{opt} (solid line), the plug-in bandwidth $\widehat{h}_{opt}$ (long dashed line) and the rule-of-thumb bandwidth h_{ROT} (short dashed line) given 100 observations. Figure 1 (b) shows the corresponding

results for 500 observations. Figure 1 (c), the last one in this column, shows the densities of the bandwidth ratios $\widehat{h}_{opt}/h_{opt}$ for 100 observations (thin, solid line) and 500 observations (thick, solid line) as well as the densities of the bandwidth ratios h_{ROT}/h_{opt} for 100 observations (thin, short dashed line) and 500 observations (thick, short dashed line). The corresponding graphs for Model 2 are Figures 1 (d) - (f), while Figure 2 (a) - (f) contains all the graphs for Model 3 and 4.

¿From inspecting all upper and middle Figures one can see that the densities of the EMISE of the plug-in bandwidth are always to the left of the EMISE densities of the rule-of-thumb bandwidth, independently of the model chosen and the sample size. Therefore, conducting the additional effort in estimating $C(f)$ by the partial local quadratic estimator pays off even in small samples of 100 observations. Furthermore, the densities of the EMISE associated with the plug-in bandwidth $\widehat{h}_{opt}$ are located extremely close to the densities associated with the optimal bandwidth h_{opt}. This indicates that the $\widehat{C}(f)$ estimate is quite close to the true $C(f)$. The EMISE of h_{opt} is the least spread out of all three due to the fact that h_{opt} is constant across all replications for a fixed model and sample size. Experimenting with other h_C, that is with smaller factors to be multiplied with $h_S(d+4)$, led to a substantial increase in the variance of the $\widehat{C}(f)$ and $\widehat{h}_{opt}$ estimates. This can be seen by comparing the scales of the Figures (a), (d) and (b), (e) for 100 and 500 observations, respectively. It is worth pointing out that these results are obtained for quite different regression functions.

The last Figure in each column allows to study the behavior of the bandwidth ratios $\widehat{h}_{opt}/h_{opt}$ and h_{ROT}/h_{opt}. First of all, one finds that for none of the four models (Figures 1(c), (f), 2(c), (f)) the center of the densities of the bandwidth ratio h_{ROT}/h_{opt} (thin or thick dashed lines) converge to 1. In contrast, as implied by Theorem 2.1 this convergence occurs for the densities of the bandwidth ratio $\widehat{h}_{opt}/h_{opt}$ (thin or thick solid lines) for all four models. However, one observes the expected decrease in variation only for Model 1 indicating the influence of the regression function on when the asymptotics kick in. Among the four models, Model 1 has the simplest regression function since it is the only one with constant second derivatives.

Overall, this Monte Carlo evidence is very supportive for the plug-in bandwidth $\widehat{h}_{opt}$ because it is possible to obtain reasonable estimates of the integrated squared Laplacian of a multivariate regression function by a suitable bandwidth h_C and a partial local quadratic estimator.

5. CONCLUSION

In this paper we proposed a plug-in bandwidth for multivariate local linear regression. This plug-in bandwidth estimates the optimal bandwidth by estimating unknown functionals, including the integrated squared Laplacian of the multivariate regression function. In a Monte Carlo study, our plug-in bandwidth was found to approximate the optimal bandwidth much better and produce much smaller mean integrated squared error than a naive rule-of-thumb bandwidth.

To estimate the integrated squared Laplacian of the regression function by a partial local quadratic estimator, we derived an optimal bandwidth via a within bias trade-off. As this optimal bandwidth includes highly complicated functionals, we suggested a feasible pilot bandwidth that retains the optimal rate. In doing so, we tried more to keep this pilot bandwidth from being too small than too large. This is due to the within bias trade-off: Unlike the usual situation

where a smaller bandwidth means a smaller bias, here deviating from the optimal bandwidth always increases bias. On the other hand, increasing the bandwidth leads to a smaller variance as usual. Our Monte Carlo results on the plug-in bandwidth indicate satisfactory performance of this pilot bandwidth despite the curse of dimensionality. We believe that the multivariate functional estimation techniques employed in this paper can be useful in other situations as well.

APPENDIX

In order to prove Theorems 3.1 and 3.2, we need the following assumptions

(A1) The density $p(\mathbf{x})$ of $\mathbf{X}$ exists and is continuously differentiable everywhere up to order two on its support S.

(A2) The function $f(\mathbf{x})$ is continuously differentiable on S up to order four while $g(\mathbf{x})$ is continuous on S.

(A3) The bandwidth $h = h_n$ is a positive number depending on n such that $h \to 0$ and $nh^{d+4} \to \infty$.

(A4) For each $\lambda = 1, 2, ..., d$, one has $\int f_{\lambda\lambda}^2(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} < \infty$. Also $\int g^4(\mathbf{x}) d\mathbf{x} < \infty$. Here we make the convention that all integrations in variable $\mathbf{x}$ are over S.

Lemma 5.1 *Let Z be as in (3.6) and*

$$H = \begin{pmatrix} 1 & 0_{1\times d} & 0_{1\times d} \\ 0_{d\times 1} & h^{-1} I_d & 0_{d\times d} \\ 0_{d\times 1} & 0_{d\times d} & h^{-2} I_d \end{pmatrix}.$$

As $n \to \infty$,

$$H^T Z^T W Z H = p(\mathbf{x}) \begin{bmatrix} 1 & 0_{1\times d} & \sigma_K^2 1_{1\times d} \\ 0_{d\times 1} & \sigma_K^2 I_d & 0_{d\times d} \\ \sigma_K^2 1_{d\times 1} & 0_{d\times d} & \sigma_K^4(\kappa-1)\left\{ I_d + \frac{1}{\kappa-1} 1_{d\times 1} 1_{1\times d} \right\} \end{bmatrix} \{1 + o_p(1)\} \tag{5.1}$$

$$\left(H^T Z^T W Z H \right)^{-1} =$$

$$\frac{1}{p(\mathbf{x})} \begin{bmatrix} (\kappa - 1 + d)/(\kappa - 1) & 0_{1\times d} & -\sigma_K^{-2}(\kappa-1)^{-1} 1_{1\times d} \\ 0_{d\times 1} & \sigma_K^{-2} I_d & 0_{d\times d} \\ -\sigma_K^{-2}(\kappa-1)^{-1} 1_{d\times 1} & 0_{d\times d} & \sigma_K^{-4}(\kappa-1)^{-1} I_d \end{bmatrix} \{1 + o_p(1)\} \tag{5.2}$$

uniformly in a compact neighborhood of x.

Proof. The elements of $H^T Z^T W Z H$ are all of the following form

$$\sum_{i=1}^{n} \frac{1}{n} K_h(\mathbf{X}_i - \mathbf{x}) \frac{(X_{i\alpha} - x_\alpha)^l}{h^l} \frac{(X_{i\beta} - x_\beta)^k}{h^k}$$

where $k, l = 0, 1, 2$ while $1 \leq \alpha, \beta \leq d$. The equation (5.1) follows by the usual array type central limit theorem, K being a symmetric compact product kernel, see, for example, Wand and Jones (1995), or Fan and Gijbels (1996). The equation (5.2) follows by applying the results of Lütkepohl (1996), equation (**1**), page 30, plus equation (**3**), page 29. Specifically, by equation (**3**), page 29 of Lütkepohl (1996), one has

$$\left\{\sigma_K^4(\kappa-1)\left(I_d + \frac{1}{\kappa-1}1_{d\times 1}1_{1\times d}\right)\right\}^{-1} = \sigma_K^{-4}(\kappa-1)^{-1}\left\{I_d - \frac{1}{(\kappa-1)\left(1+\frac{d}{\kappa-1}\right)}1_{d\times 1}1_{1\times d}\right\}$$

$$= \sigma_K^{-4}(\kappa-1)^{-1}\left(I_d - \frac{1}{\kappa-1+d}1_{d\times 1}1_{1\times d}\right).$$

One then notes that by Lütkepohl (1996), equation (**1**), page 30, one has

$$\begin{bmatrix} 1 & 0_{1\times d} & \sigma_K^2 1_{1\times d} \\ 0_{d\times 1} & \sigma_K^2 I_d & 0_{d\times d} \\ \sigma_K^2 1_{d\times 1} & 0_{d\times d} & \sigma_K^4(\kappa-1)\left\{I_d + \frac{1}{\kappa-1}1_{d\times 1}1_{1\times d}\right\} \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ B' & D & 0_{d\times d} \\ C' & 0_{d\times d} & E \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} F & -FBD^{-1} & -FCE^{-1} \\ -D^{-1}B'F & D^{-1}+D^{-1}B'FBD^{-1} & D^{-1}B'FCE^{-1} \\ -E^{-1}C'F & E^{-1}C'FBD^{-1} & E^{-1}+E^{-1}C'FCE^{-1} \end{bmatrix} \tag{5.3}$$

where $F = (A - BD^{-1}B' - CE^{-1}C')^{-1}$. Now because $A = 1$, $B = 0_{1\times d}$, $C = \sigma_K^2 1_{1\times d}$, $D = \sigma_K^2 I_d$, $E = \sigma_K^4(\kappa-1)\left(I_d + \frac{1}{\kappa-1}1_{d\times 1}1_{1\times d}\right)$, we obtain

$$A - BD^{-1}B' - CE^{-1}C' = 1 - \sigma_K^2 1_{1\times d} \times \sigma_K^{-4}(\kappa-1)^{-1}\left(I_d - \frac{1}{\kappa-1+d}1_{d\times 1}1_{1\times d}\right) \times \sigma_K^2 1_{d\times 1}$$

$$= 1 - (\kappa-1)^{-1}\left(d - \frac{d^2}{\kappa-1+d}\right) = \frac{\kappa-1}{\kappa-1+d}.$$

Hence $F = (\kappa-1+d)/(\kappa-1)$. So

$$FCE^{-1} = (\kappa-1+d)/(\kappa-1) \times \sigma_K^2 1_{1\times d} \times \sigma_K^{-4}(\kappa-1)^{-1}\left(I_d - \frac{1}{\kappa-1+d}1_{d\times 1}1_{1\times d}\right)$$

$$= \sigma_K^{-2}(\kappa-1)^{-1}(\kappa-1+d)/(\kappa-1)\left(1_{1\times d} - \frac{d}{\kappa-1+d}1_{1\times d}\right) = \sigma_K^{-2}(\kappa-1)^{-1}1_{1\times d} \tag{5.4}$$

and

$$E^{-1} + E^{-1}C'FCE^{-1} = \sigma_K^{-4}(\kappa-1)^{-1}\left(I_d - \frac{1}{\kappa-1+d}1_{d\times 1}1_{1\times d}\right) + F^{-1}\left(FCE^{-1}\right)'FCE^{-1}$$

$$= \sigma_K^{-4}(\kappa-1)^{-1}\left(I_d - \frac{1}{\kappa-1+d}1_{d\times 1}1_{1\times d}\right) + \frac{(\kappa-1)}{\kappa-1+d}\sigma_K^{-4}(\kappa-1)^{-2}1_{d\times 1}1_{1\times d} = \sigma_K^{-4}(\kappa-1)^{-1}I_d. \tag{5.5}$$

Plugging equations (5.5), (5.4), $B = 0_{1\times d}$, $D = \sigma_K^2 I_d$ and $F = (\kappa - 1 + d)/(\kappa - 1)$ into (5.3) gives equation (5.2). Q.E.D.

Proof of Theorem 3.1.

Now notice that

$$\widehat{f_{\lambda\lambda}}(\mathbf{x}) - f_{\lambda\lambda}(\mathbf{x}) = 2e_{d+\lambda}^T H \left(H^T Z^T W Z H\right)^{-1} H^T Z^T W Y - f_{\lambda\lambda}(\mathbf{x})$$

and by equation (5.2) of Lemma 5.1, it becomes

$$= \frac{2}{\sigma_K^4(\kappa-1)p(\mathbf{x})nh^2}\{1+o_p(1)\}\sum_{i=1}^n K_h(X_i-\mathbf{x})\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}\{f(\mathbf{X}_i)-f_{\lambda\lambda}(\mathbf{x})+g(\mathbf{X}_i)\epsilon_i\}$$

and that

$$e_{d+\lambda}^T \left(Z^T W Z\right)^{-1} Z^T W Z e_{\lambda'} = 0$$

for any $\lambda' \neq d+\lambda$, one can add in terms like

$$e_{d+\lambda}^T \left(Z^T W Z\right)^{-1} Z^T W Z e_0 f(\mathbf{x}) = e_{d+\lambda}^T H \left(H^T Z^T W Z H\right)^{-1} H^T Z^T W Z e_0 f(\mathbf{x})$$

or

$$e_{d+\lambda}^T \left(Z^T W Z\right)^{-1} Z^T W Z e_{\lambda'} f_{\lambda'}(\mathbf{x}) = e_{d+\lambda}^T H \left(H^T Z^T W Z H\right)^{-1} H^T Z^T W Z e_{\lambda'} f_{\lambda'}(\mathbf{x})$$

where $\lambda' = 1, ..., d$ or

$$e_{d+\lambda}^T H \left(H^T Z^T W Z H\right)^{-1} H^T Z^T W Z e_{d+\lambda'} f_{\lambda'\lambda'}(\mathbf{x})$$

where $\lambda' = 1, ..., d$ and $\lambda' \neq \lambda$. Thus $\widehat{f_{\lambda\lambda}}(\mathbf{x}) - f_{\lambda\lambda}(\mathbf{x})$ becomes

$$T_1 + T_2 + T_3$$

where

$$T_1 = \frac{1}{\sigma_K^4(\kappa-1)p(\mathbf{x})nh^2}\{1+o_p(1)\}\sum_{i=1}^n K_h(\mathbf{X}_i-\mathbf{x})\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}$$
$$\left\{f(\mathbf{X}_i)-f(\mathbf{x})-(\mathbf{X}_i-\mathbf{x})^T\nabla f(\mathbf{x})-\frac{1}{2}(\mathbf{X}_i-\mathbf{x})^T\nabla^2 f(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\right\}$$

$$T_2 = -\frac{\{1+o_p(1)\}}{\sigma_K^4(\kappa-1)p(\mathbf{x})nh^2}\sum_{i=1}^n K_h(\mathbf{X}_i-\mathbf{x})\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}\left\{\sum_{1\le\alpha<\beta\le d}(X_{i\alpha}-x_\alpha)(X_{i\beta}-x_\beta)f_{\alpha\beta}(\mathbf{x})\right\}$$

$$T_3 = \frac{1}{\sigma_K^4(\kappa-1)p(\mathbf{x})nh^2}\{1+o_p(1)\}\sum_{i=1}^n K_h(\mathbf{X}_i-\mathbf{x})\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}g(\mathbf{X}_i)\epsilon_i$$

where the asymptotic variance of T_3 is calculated as

$$\frac{1}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})^2nh^4}\{1+o(1)\}\int K_h(\mathbf{u}-\mathbf{x})^2\frac{(u_\lambda-x_\lambda)^4}{h^4}g^2(\mathbf{u})p(\mathbf{u})d\mathbf{u}$$

which is (using $\mathbf{u} = \mathbf{x} + h\mathbf{v}$)

$$= \frac{1}{\sigma_K^8(\kappa-1)^2 p(\mathbf{x})^2 n h^{d+4}} \{1+o(1)\} \int K(\mathbf{v})^2 v_\lambda^4 g^2(\mathbf{x}+h\mathbf{v}) p(\mathbf{x}+h\mathbf{v}) d\mathbf{v}$$

$$= \frac{\mu_4(K^2) \|K\|_2^{2(d-1)} g^2(\mathbf{x})}{\sigma_K^8(\kappa-1)^2 p(\mathbf{x}) n h^{d+4}} \{1+o(1)\}.$$

While the similar procedure applied to T_1 yields

$$T_1 = \frac{1}{\sigma_K^4(\kappa-1)p(\mathbf{x})h^2} \{1+o_p(1)\} \int K_h(\mathbf{u}-\mathbf{x}) \frac{(u_\lambda - x_\lambda)^2}{h^2}$$

$$\left\{ f(\mathbf{u}) - f(\mathbf{x}) - (\mathbf{u}-\mathbf{x})^T \nabla f(\mathbf{x}) - \frac{1}{2}(\mathbf{u}-\mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{u}-\mathbf{x}) \right\} p(\mathbf{u}) d\mathbf{u}$$

$$= \frac{\{1+o_p(1)\}}{\sigma_K^4(\kappa-1)p(\mathbf{x})h^2} \int K(\mathbf{v}) v_\lambda^2 \left\{ f(\mathbf{x}+h\mathbf{v}) - f(\mathbf{x}) - h\mathbf{v}^T \nabla f(\mathbf{x}) - \frac{1}{2} h^2 \mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \right\} p(\mathbf{x}+h\mathbf{v}) d\mathbf{v}$$

$$= \frac{h^2}{\sigma_K^4(\kappa-1)p(\mathbf{x})} \left[\sum_{1 \le \alpha < \beta \le d} \left\{ \frac{1}{2} f_{\alpha\alpha\beta}(\mathbf{x}) p_\beta(\mathbf{x}) + \frac{1}{2} f_{\alpha\beta\beta}(\mathbf{x}) p_\alpha(\mathbf{x}) + \frac{1}{4} f_{\alpha\alpha\beta\beta}(\mathbf{x}) p(\mathbf{x}) \right\} \right] \int K(\mathbf{v}) v_\lambda^2 v_\alpha^2 v_\beta^2 d\mathbf{v}$$

$$+ \frac{h^2}{\sigma_K^4(\kappa-1)p(\mathbf{x})} \left[\sum_{\gamma=1}^{d} \left\{ \frac{1}{6} f_{\gamma\gamma\gamma}(\mathbf{x}) p_\gamma(\mathbf{x}) + \frac{1}{24} f_{\gamma\gamma\gamma\gamma}(\mathbf{x}) p(\mathbf{x}) \right\} \right] \int K(\mathbf{v}) v_\lambda^2 v_\gamma^4 d\mathbf{v} + o(h^2)$$

The term T_2 is treated as

$$T_2 = -\frac{\{1+o_p(1)\}}{\sigma_K^4(\kappa-1)p(\mathbf{x})n} \sum_{i=1}^{n} K_h(\mathbf{X}_i - \mathbf{x}) \frac{(X_{i\lambda} - x_\lambda)^2}{h^2} \left\{ \sum_{1 \le \alpha < \beta \le d} \left(\frac{X_{i\alpha} - x_\alpha}{h} \right) \left(\frac{X_{i\beta} - x_\beta}{h} \right) f_{\alpha\beta}(\mathbf{x}) \right\}$$

$$= -\sum_{1 \le \alpha < \beta \le d} \frac{f_{\alpha\beta}(\mathbf{x})}{\sigma_K^4(\kappa-1)p(\mathbf{x})} \{1+o_p(1)\} \int K_h(\mathbf{u}-\mathbf{x}) \frac{(u_\lambda - x_\lambda)^2}{h^2} \left(\frac{u_\alpha - x_\alpha}{h} \right) \left(\frac{u_\beta - x_\beta}{h} \right) p(\mathbf{u}) d\mathbf{u}$$

$$= -\sum_{1 \le \alpha < \beta \le d} \frac{f_{\alpha\beta}(\mathbf{x})}{\sigma_K^4(\kappa-1)p(\mathbf{x})} \int K(\mathbf{v}) v_\lambda^2 v_\alpha v_\beta p(\mathbf{x}+h\mathbf{v}) d\mathbf{v} \{1+o_p(1)\}$$

$$= -\frac{h^2}{2} \sum_{1 \le \alpha < \beta \le d} \frac{f_{\alpha\beta}(\mathbf{x})}{\sigma_K^4(\kappa-1)p(\mathbf{x})} \int K(\mathbf{v}) v_\lambda^2 v_\alpha^2 v_\beta^2 p_{\alpha\beta}(\mathbf{x}) d\mathbf{v} \{1+o_p(1)\}.$$

Now adding T_1 and T_2 gives

$$T_1 + T_2 = b_{\lambda\lambda}(\mathbf{x}) h^2$$

in which the bias coefficient is

$$b_{\lambda\lambda}(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \left[\frac{\kappa}{\kappa-1} \sigma_K^2 \sum_{\gamma \ne \lambda} \left\{ \frac{f_{\gamma\gamma\gamma}(\mathbf{x}) p_\gamma(\mathbf{x})}{6} + \frac{f_{\gamma\gamma\gamma\gamma}(\mathbf{x}) p(\mathbf{x})}{24} \right\} \right]$$

$$+\frac{1}{p(\mathbf{x})}\left[\frac{\mu_6(K)}{\sigma_K^4(\kappa-1)}\left\{\frac{f_{\lambda\lambda\lambda}(\mathbf{x})p_\lambda(\mathbf{x})}{6}+\frac{f_{\lambda\lambda\lambda\lambda}(\mathbf{x})p(\mathbf{x})}{24}\right\}\right]$$

$$+\frac{1}{p(\mathbf{x})}\left[\frac{\sigma_K^2}{\kappa-1}\sum_{\alpha<\beta,\alpha,\beta\neq\lambda}\left\{\frac{f_{\alpha\alpha\beta}(\mathbf{x})p_\beta(\mathbf{x})}{2}+\frac{f_{\alpha\beta\beta}(\mathbf{x})p_\alpha(\mathbf{x})}{2}+\frac{f_{\alpha\alpha\beta\beta}(\mathbf{x})p(\mathbf{x})}{4}-\frac{f_{\alpha\beta}(\mathbf{x})p_{\alpha\beta}(\mathbf{x})}{2}\right\}\right]$$

$$+\frac{1}{p(\mathbf{x})}\left[\frac{\kappa}{\kappa-1}\sigma_K^2\sum_{\lambda<\beta}\left\{\frac{f_{\lambda\lambda\beta}(\mathbf{x})p_\beta(\mathbf{x})}{2}+\frac{f_{\lambda\beta\beta}(\mathbf{x})p_\lambda(\mathbf{x})}{2}+\frac{f_{\lambda\lambda\beta\beta}(\mathbf{x})p(\mathbf{x})}{4}-\frac{f_{\lambda\beta}(\mathbf{x})p_{\lambda\beta}(\mathbf{x})}{2}\right\}\right]$$

$$+\frac{1}{p(\mathbf{x})}\left[\frac{\kappa}{\kappa-1}\sigma_K^2\sum_{\alpha<\lambda}\left\{\frac{f_{\alpha\alpha\lambda}(\mathbf{x})p_\lambda(\mathbf{x})}{2}+\frac{f_{\alpha\lambda\lambda}(\mathbf{x})p_\alpha(\mathbf{x})}{2}+\frac{f_{\alpha\alpha\lambda\lambda}(\mathbf{x})p(\mathbf{x})}{4}-\frac{f_{\alpha\lambda}(\mathbf{x})p_{\alpha\lambda}(\mathbf{x})}{2}\right\}\right] \quad (5.6)$$

In the mean time, the term T_3 gives the variance formula for $\sigma_{\lambda\lambda}(\mathbf{x})$ as in (3.9). Q. E. D.

Proof of Theorem 3.2.

Since

$$\widehat{f_{\lambda\lambda}}(\mathbf{x}) = f_{\lambda\lambda}(\mathbf{x}) + \left\{b_{\lambda\lambda}(\mathbf{x})h^2 + T_{3,\lambda\lambda}\right\}\left\{1+o_p\left(h^2+\frac{1}{\sqrt{nh^{d+4}}}\right)\right\}$$

where

$$T_{3,\lambda\lambda} = \frac{1}{\sigma_K^4(\kappa-1)p(\mathbf{x})nh^2}\sum_{i=1}^n K_h(\mathbf{X}_i-\mathbf{x})\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}g(\mathbf{X}_i)\epsilon_i$$

one has

$$\widehat{C}(f,h) = \int\left\{\sum_{\lambda=1}^d \widehat{f}_{\lambda\lambda}(\mathbf{x})\right\}^2 p(\mathbf{x})d\mathbf{x}$$

$$= \int\left[\sum_{\lambda=1}^d f_{\lambda\lambda}(\mathbf{x}) + \sum_{\lambda=1}^d\left\{b_{\lambda\lambda}(\mathbf{x})h^2+T_{3,\lambda\lambda}\right\}\left\{1+o_p\left(h^2+\frac{1}{\sqrt{nh^{d+4}}}\right)\right\}\right]^2 p(\mathbf{x})d\mathbf{x}$$

$$= C(f) + 2h^2\int\sum_{\lambda=1}^d f_{\lambda\lambda}(\mathbf{x})\sum_{\lambda=1}^d b_{\lambda\lambda}(\mathbf{x})p(\mathbf{x})d\mathbf{x} + o_p\left(h^4+\frac{h^2}{\sqrt{nh^{d+4}}}\right)$$

$$+\int\left(\sum_{\lambda=1}^d T_{3,\lambda\lambda}\right)^2 p(\mathbf{x})d\mathbf{x}\left\{1+o_p\left(h^2+\frac{1}{\sqrt{nh^{d+4}}}\right)\right\}. \quad (5.7)$$

Thus it remains to compute the term $\int\left(\sum_{\lambda=1}^d T_{3,\lambda\lambda}\right)^2 p(\mathbf{x})d\mathbf{x}$, which has a simple decomposition as

$$\sum_{1\le i<j\le n} 2H(\mathbf{X}_i,\mathbf{X}_j)\epsilon_i\epsilon_j + \sum_{1\le i\le n} H(\mathbf{X}_i,\mathbf{X}_i)\epsilon_i^2$$

where

$$H(\mathbf{X}_i,\mathbf{X}_j) =$$

$$\int\frac{1}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})n^2h^4}K_h(\mathbf{X}_i-\mathbf{x})\sum_{\lambda=1}^d\frac{(X_{i\lambda}-x_\lambda)^2}{h^2}g(\mathbf{X}_i)K_h(\mathbf{X}_j-\mathbf{x})\sum_{\mu=1}^d\frac{(X_{j\mu}-x_\mu)^2}{h^2}g(\mathbf{X}_j)d\mathbf{x}.$$

Note first that

$$E\left\{H(\mathbf{X}_1,\mathbf{X}_1)\epsilon_1^2\right\} = E\int \frac{d\mathbf{x}}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})n^2h^4}K_h^2(\mathbf{X}_1-\mathbf{x})\left\{\sum_{\lambda=1}^d \frac{(X_{1\lambda}-x_\lambda)^2}{h^2}\right\}^2 g^2(\mathbf{X}_1)$$

$$= \int\int \frac{d\mathbf{x}}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})n^2h^4}K_h^2(\mathbf{y}-\mathbf{x})\left\{\sum_{\lambda=1}^d \frac{(y_\lambda-x_\lambda)^2}{h^2}\right\}^2 g^2(\mathbf{y})p(\mathbf{y})d\mathbf{y}$$

$$\stackrel{\mathbf{y}-\mathbf{x}=h\mathbf{u}}{=} \int\int \frac{1}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})n^2h^{d+4}}K^2(\mathbf{u})\left(\sum_{\lambda=1}^d u_\lambda^2\right)^2 g^2(\mathbf{x}+h\mathbf{u})p(\mathbf{x}+h\mathbf{u})d\mathbf{x}d\mathbf{u}$$

$$= \frac{1}{\sigma_K^8(\kappa-1)^2n^2h^{d+4}}\int g^2(\mathbf{x})d\mathbf{x}\int K^2(\mathbf{u})\left(\sum_{\lambda=1}^d u_\lambda^2\right)^2 d\mathbf{u}\left\{1+o(1)\right\}$$

$$= \frac{\int g^2(\mathbf{x})d\mathbf{x}\left\{d\mu_4(K^2)\left\|K\right\|_2^{2(d-1)} + d(d-1)\mu_2^2(K^2)\left\|K\right\|_2^{2(d-2)}\right\}}{\sigma_K^8(\kappa-1)^2n^2h^{d+4}}\left\{1+o(1)\right\}$$

and similarly that

$$E\left\{H(\mathbf{X}_1,\mathbf{X}_1)^2\epsilon_1^4\right\} = O\left(\frac{1}{n^4h^{8+3d}}\right).$$

Hence by an array type central limit theorem

$$\sum_{1\le i\le n} H(\mathbf{X}_i,\mathbf{X}_i)\epsilon_i^2 = \frac{\int g^2(\mathbf{x})d\mathbf{x}\left\{d\mu_4(K^2)\left\|K\right\|_2^{2(d-1)} + d(d-1)\mu_2^2(K^2)\left\|K\right\|_2^{2(d-2)}\right\}}{\sigma_K^8(\kappa-1)^2nh^{d+4}}.$$

$$+O_p\left(\frac{1}{\sqrt{n^3h^{8+3d}}}\right). \tag{5.8}$$

Meanwhile, using a central limit theorem for nondegenerate *U-statistics* as contained in Hall (1984), one can verify that $\sum_{1\le i<j\le n} 2H(\mathbf{X}_i,\mathbf{X}_j)\epsilon_i\epsilon_j$ is asymptotically normal with mean zero and asymptotic variance

$$\frac{n^2}{2}E\left\{4H(\mathbf{X}_1,\mathbf{X}_2)^2\epsilon_1^2\epsilon_2^2\right\} = 2n^2E\left\{H(\mathbf{X}_1,\mathbf{X}_2)^2\right\} =$$

$$2n^2E\left\{\int \frac{1}{\sigma_K^8(\kappa-1)^2p(\mathbf{x})n^2h^4}K_h(\mathbf{X}_1-\mathbf{x})\sum_{\lambda=1}^d \frac{(X_{1\lambda}-x_\lambda)^2}{h^2}g(\mathbf{X}_1)K_h(\mathbf{X}_2-\mathbf{x})\sum_{\mu=1}^d \frac{(X_{2\mu}-x_\mu)^2}{h^2}g(\mathbf{X}_2)d\mathbf{x}\right\}^2$$

which is

$$\frac{2}{\sigma_K^{16}(\kappa-1)^4n^2h^8}E\left\{\int \frac{1}{p(\mathbf{x})}K_h(\mathbf{X}_1-\mathbf{x})\sum_{\lambda=1}^d \frac{(X_{1\lambda}-x_\lambda)^2}{h^2}g(\mathbf{X}_1)K_h(\mathbf{X}_2-\mathbf{x})\sum_{\mu=1}^d \frac{(X_{2\mu}-x_\mu)^2}{h^2}g(\mathbf{X}_2)d\mathbf{x}\right\}^2$$

$$= \frac{2}{\sigma_K^{16}(\kappa-1)^4n^2h^8}\int\left\{\int \frac{K_h(\mathbf{y}_1-\mathbf{x})K_h(\mathbf{y}_2-\mathbf{x})}{p(\mathbf{x})}\sum_{\lambda=1}^d \frac{(y_{1\lambda}-x_\lambda)^2}{h^2}\sum_{\mu=1}^d \frac{(y_{2\mu}-x_\mu)^2}{h^2}g(\mathbf{y}_1)g(\mathbf{y}_2)d\mathbf{x}\right\}^2$$

$$\times p(\mathbf{y}_1)p(\mathbf{y}_2)d\mathbf{y}_1 d\mathbf{y}_2$$

$$= \frac{2}{\sigma_K^{16}(\kappa-1)^4 n^2 h^8} \int \frac{K_h(\mathbf{y}_1-\mathbf{x}_1)K_h(\mathbf{y}_2-\mathbf{x}_1)K_h(\mathbf{y}_1-\mathbf{x}_2)K_h(\mathbf{y}_2-\mathbf{x}_2)}{p(\mathbf{x}_1)p(\mathbf{x}_2)}$$

$$\times \sum_{\lambda=1}^{d} \frac{(y_{1\lambda}-x_{1\lambda})^2}{h^2} \sum_{\mu=1}^{d} \frac{(y_{2\mu}-x_{1\mu})^2}{h^2} \sum_{\lambda=1}^{d} \frac{(y_{1\lambda}-x_{2\lambda})^2}{h^2} \sum_{\mu=1}^{d} \frac{(y_{2\mu}-x_{2\mu})^2}{h^2}$$

$$\times g^2(\mathbf{y}_1)g^2(\mathbf{y}_2)p(\mathbf{y}_1)p(\mathbf{y}_2)d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{y}_1 d\mathbf{y}_2.$$

This, after doing a change of variables $\mathbf{y}_1 - \mathbf{x}_1 = h\mathbf{u}_1$, $\mathbf{y}_2 - \mathbf{x}_2 = h\mathbf{u}_2$, becomes

$$= \frac{2}{\sigma_K^{16}(\kappa-1)^4 n^2 h^8} \int\int \frac{K(\mathbf{u}_1)K_h(\mathbf{x}_2-\mathbf{x}_1+h\mathbf{u}_2)K_h(\mathbf{x}_1-\mathbf{x}_2+h\mathbf{u}_1)K(\mathbf{u}_2)}{p(\mathbf{x}_1)p(\mathbf{x}_2)}$$

$$\times \sum_{\lambda=1}^{d} u_{1\lambda}^2 \sum_{\mu=1}^{d} \frac{(x_{2\mu}-x_{1\mu}+hu_{2\mu})^2}{h^2} \sum_{\lambda=1}^{d} \frac{(x_{1\lambda}-x_{2\lambda}+hu_{1\lambda})^2}{h^2} \sum_{\mu=1}^{d} u_{2\mu}^2$$

$$\times g^2(\mathbf{x}_1+h\mathbf{u}_1)g^2(\mathbf{x}_2+h\mathbf{u}_2)p(\mathbf{x}_1+h\mathbf{u}_1)p(\mathbf{x}_2+h\mathbf{u}_2)d\mathbf{x}_1 d\mathbf{x}_2 d\mathbf{u}_1 d\mathbf{u}_2$$

using another change of variables $\mathbf{x}_2 - \mathbf{x}_1 = h\mathbf{t}_1$, the above becomes

$$= \frac{2}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int\int \frac{K(\mathbf{u}_1)K(\mathbf{t}_1+\mathbf{u}_2)K(-\mathbf{t}_1+\mathbf{u}_1)K(\mathbf{u}_2)}{p(\mathbf{x}_1)p(\mathbf{x}_1+h\mathbf{t}_1)}$$

$$\times \sum_{\lambda=1}^{d} u_{1\lambda}^2 \sum_{\mu=1}^{d} (t_{1\mu}+u_{2\mu})^2 \sum_{\lambda=1}^{d} (-t_{1\lambda}+u_{1\lambda})^2 \sum_{\mu=1}^{d} u_{2\mu}^2$$

$$\times g^2(\mathbf{x}_1+h\mathbf{u}_1)g^2(\mathbf{x}_1+h\mathbf{t}_1+h\mathbf{u}_2)p(\mathbf{x}_1+h\mathbf{u}_1)p(\mathbf{x}_1+h\mathbf{t}_1+h\mathbf{u}_2)d\mathbf{x}_1 d\mathbf{t}_1 d\mathbf{u}_1 d\mathbf{u}_2$$

$$= \frac{2\int g^4(\mathbf{x}_1)d\mathbf{x}_1}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int K(\mathbf{u}_1) \sum_{\lambda=1}^{d} u_{1\lambda}^2 K(-\mathbf{t}_1+\mathbf{u}_1) \sum_{\lambda=1}^{d} (-t_{1\lambda}+u_{1\lambda})^2$$

$$K(\mathbf{u}_2) \sum_{\mu=1}^{d} u_{2\mu}^2 K(\mathbf{t}_1+\mathbf{u}_2) \sum_{\mu=1}^{d} (t_{1\mu}+u_{2\mu})^2 d\mathbf{t}_1 d\mathbf{u}_1 d\mathbf{u}_2 \{1+o(1)\}$$

$$= \frac{2\int g^4(\mathbf{x}_1)d\mathbf{x}_1}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int d\mathbf{t}_1 \left\{ \int K(\mathbf{t}_1+\mathbf{u}_2)K(\mathbf{u}_2) \sum_{\mu=1}^{d} (t_{1\mu}+u_{2\mu})^2 \sum_{\lambda=1}^{d} u_{2\lambda}^2 d\mathbf{u}_2 \right\}^2 \{1+o(1)\}$$

$$= \frac{2\int g^4(\mathbf{x}_1)d\mathbf{x}_1}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int d\mathbf{t}_1 F^2(\mathbf{t}_1) \{1+o(1)\}$$

where

$$F(\mathbf{t}_1) = \sum_{1\le\lambda\ne\mu\le d} (K^* * K)(t_{1\mu}) ((K^* * K))(t_{1\lambda}) \left\{ \prod_{\gamma\ne\lambda,\mu} K^{(2)}(t_{1\gamma}) \right\} + \sum_{1\le\lambda\le d} K^{*(2)}(t_{1\lambda}) \left\{ \prod_{\gamma\ne\lambda} K^{(2)}(t_{1\gamma}) \right\}.$$

Hence $\zeta = \sum_{1 \le i < j \le n} 2H(\mathbf{X}_i, \mathbf{X}_j)\epsilon_i\epsilon_j$ has mean zero and asymptotic variance

$$\frac{2\int g^4(\mathbf{x})d\mathbf{x}}{\sigma_K^{16}(\kappa-1)^4 n^2 h^{d+8}} \int d\mathbf{t} F^2(\mathbf{t}).$$

This, plus equations (5.7) and (5.8) have completed the proof of Theorem 3.2. Q. E. D.

References

[1] Cheng, M. Y. (1997), A Bandwidth Selector for Local Linear Density Estimators, *Annals of Statistics*, 25, 1001-1013.

[2] Fan, J. and Gijbels, I. (1995), Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction, *Journal of Computational and Graphical Statistics*, 4, 213-227.

[3] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

[4] Grund, B. and Polzehl, J. (1997), Bias Corrected Bootstrap Bandwidth Selection, *Journal of Nonparametric Statistics*, to appear.

[5] Hall, P. (1984), Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, *Journal of Multivariate Analysis*, 14, 1-16.

[6] Härdle, W., Tsybakov, A. B. and Yang, L. (1997), Nonparametric Vector Autoregression, *Journal of Statistical Planning and Inference*, to appear.

[7] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a), A Brief Survey of Bandwidth Selection for Density Estimation, *Journal of the American Statistical Association*, 91, 401-407.

[8] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b), Progress in Data-Based Bandwidth Selection for Kernel Density Estimation, *Computational Statistics*, 11, 337-381.

[9] Lütkepohl, H. (1996), *Handbook of Matrices*, John Wiley and Sons, Chichester.

[10] Neumann, M. (1995), Automatic Bandwidth Choice and Confidence Intervals in Nonparametric Regression, *Annals of Statistics*, 23, 1937 -1959.

[11] Ruppert, D. (1997), Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation, *Journal of the American Statistical Association*, 92, 1049-1062.

[12] Ruppert, D. and Wand, M. P. (1994), Multivariate Locally Weighted Least Squares Regression, *Annals of Statistics*, 22, 1346-1370.

[13] Ruppert, D. Sheather, S. J. and Wand, M. P. (1995), An Effective Bandwidth Selector for Local Least Squares Regression, *Journal of the American Statistical Association*, 90, 1257-1270.

[14] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[15] Tjøstheim, D. and Auestad, B. (1994), Nonparametric Identification of Nonlinear Time Series: Selecting Significant Lags, *Journal of the American Statistical Association*, 89, 1410-1419.

[16] Tschernig, R. and Yang, L. (1997), Nonparametric Lag Selection for Time Series, Discussion Paper 59, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

[17] Wand, M. P. and Jones, M. C. (1993), Comparison of Smoothing Parametrizations in Bivariate Kernel Density Estimation, *Journal of the American Statistical Association*, 88, 520-528.

[18] Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman and Hall, London.

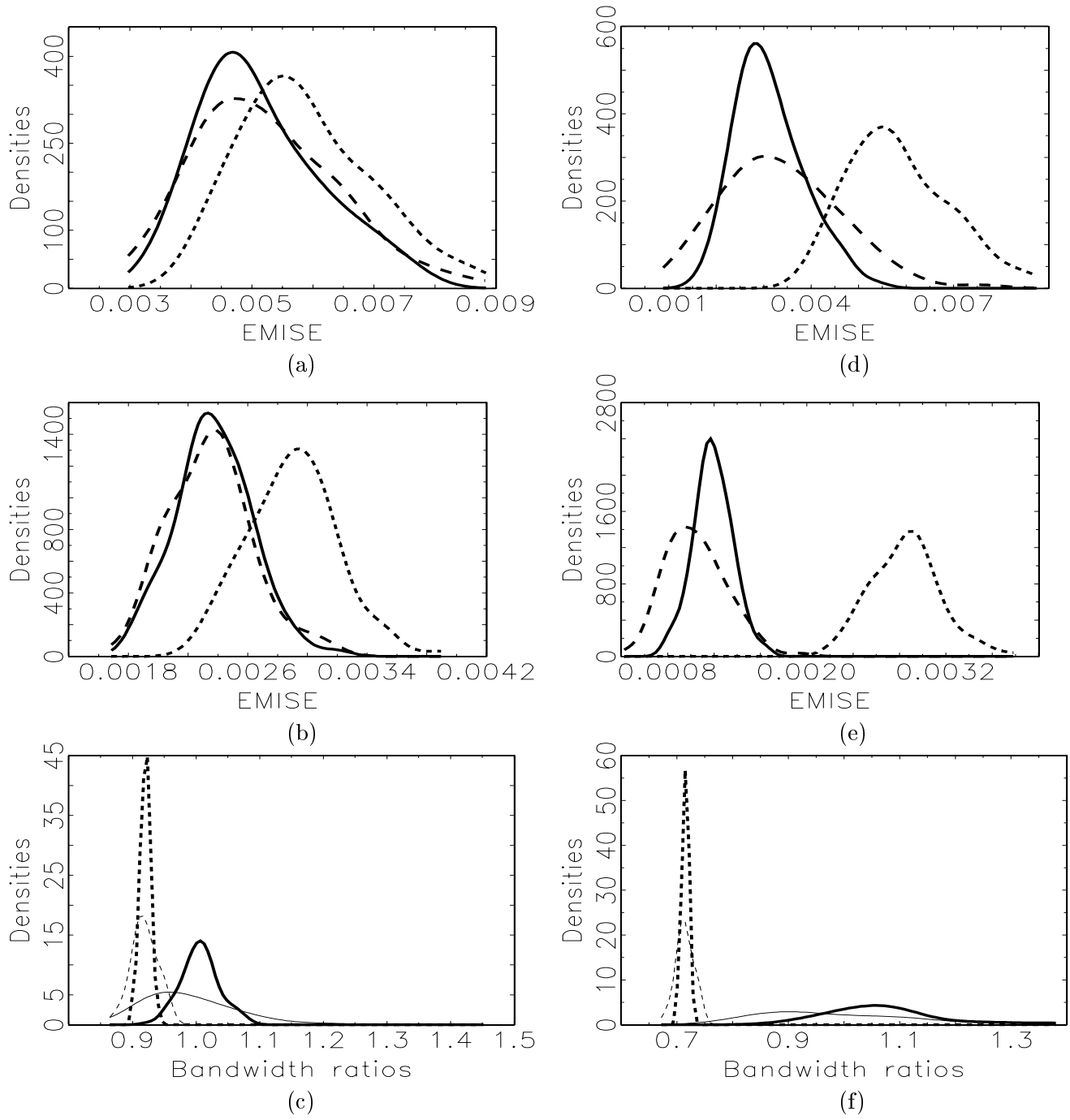

Figure 1: Model 1: (a) Estimated mean integrated squared error, 100 observations: solid line - optimal bandwidth; long dashed line - Plug-in bandwidth; short dashed line - ROT; (b) Estimated mean integrated squared error, 500 observations: solid line - optimal bandwidth; long dashed line - Plug-in bandwidth; short dashed line - ROT; (c) Ratio of Plug-in and optimal bandwidth (solid line) and ratio of ROT and optimal bandwidth (short dashes) with 100 observations (thin lines) and 500 observations (thick lines); (d), (e) and (f) are for Model 2, respectively.

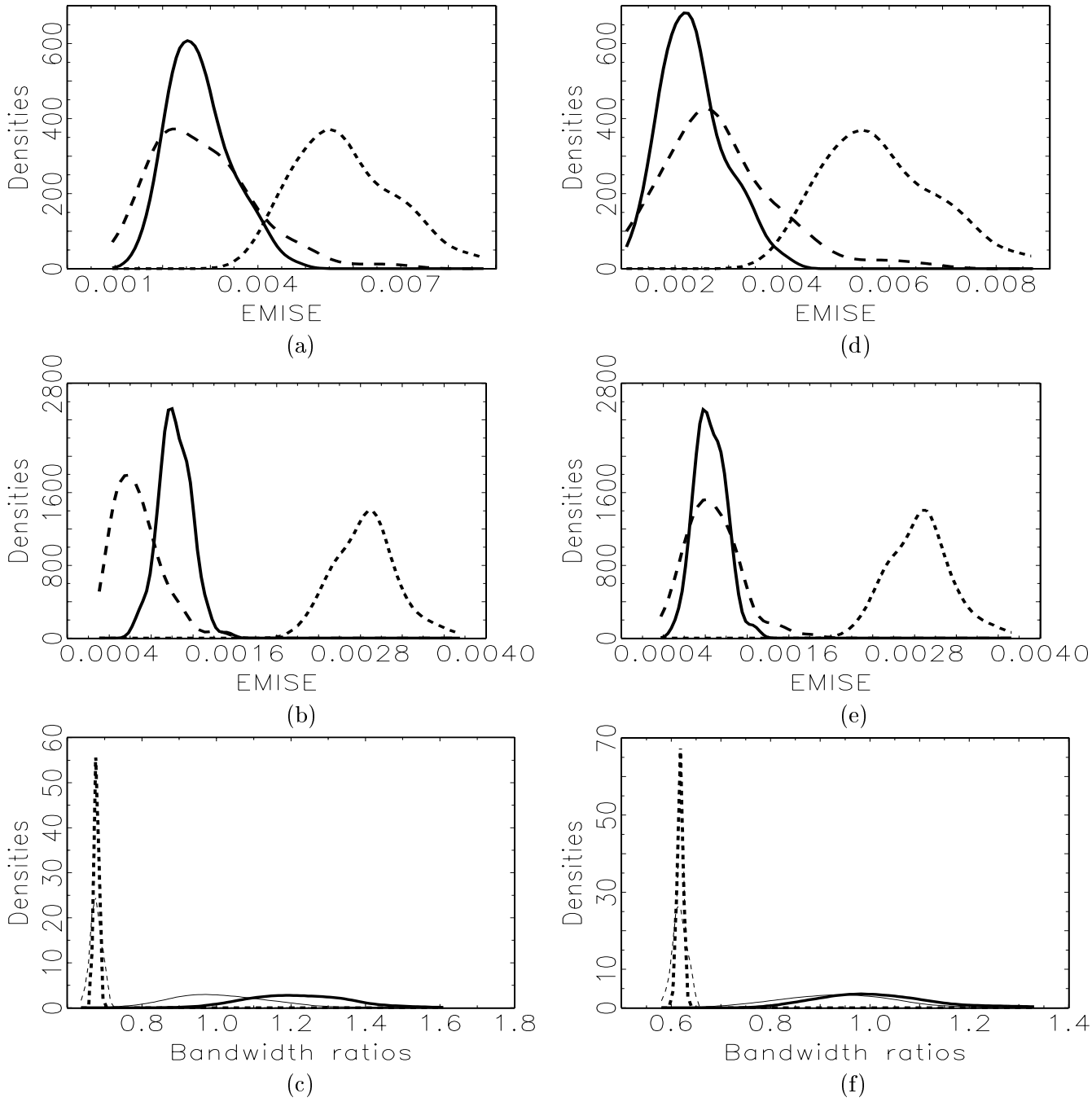

Figure 2: Model 3: (a), (b) and (c); Model 4: (d), (e) and (f); see Figure 1 for details.